Estimation of the Error Density in a Semiparametric Transformation Model

Benjamin Colling
Université catholique de Louvain *

Cédric Heuchenne
University of Liège and Université catholique de Louvain *

Rawane Samb
Université catholique de Louvain ‡

Ingrid Van Keilegom
Université catholique de Louvain §

November 14, 2013

Abstract

Consider the semiparametric transformation model $\Lambda_{\theta_o}(Y) = m(X) + \varepsilon$, where θ_o is an unknown finite dimensional parameter, the functions Λ_{θ_o} and m are smooth, ε is independent of X, and $\mathbb{E}(\varepsilon) = 0$. We propose a kernel-type estimator of the density of the error ε, and prove its asymptotic normality. The estimated errors, which lie at the basis of this estimator, are obtained from a profile likelihood estimator of θ_o and a nonparametric kernel estimator of m. The practical performance of the proposed density estimator is evaluated in a simulation study.

Key Words: Density estimation; Kernel smoothing; Nonparametric regression; Profile likelihood; Transformation model.

Running Head: Error density estimation in transformation models

*Research supported by IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

‡Research supported by IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

§Research supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, by IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent replicates of the random vector (X, Y), where Y is a univariate dependent variable and X is a one-dimensional covariate. We assume that Y and X are related via the semiparametric transformation model

$$\Lambda_{\theta_o}(Y) = m(X) + \varepsilon, \tag{1}$$

where ε is independent of X and has mean zero. We assume that $\{\Lambda_\theta : \theta \in \Theta\}$ (with $\Theta \subset \mathbb{R}^p$ compact) is a parametric family of strictly increasing functions defined on an unbounded subset $\mathcal{D}$ in $\mathbb{R}$, while m is the unknown regression function, belonging to an infinite dimensional parameter set $\mathcal{M}$. We assume that $\mathcal{M}$ is a space of functions endowed with the norm $\|\cdot\|_{\mathcal{M}} = \|\cdot\|_\infty$. We denote $\theta_o \in \Theta$ and $m \in \mathcal{M}$ for the true unknown finite and infinite dimensional parameters. Define the regression function

$$m_\theta(x) = \mathbb{E}[\Lambda_\theta(Y)|X = x],$$

for each $\theta \in \Theta$, and let $\varepsilon_\theta = \varepsilon(\theta) = \Lambda_\theta(Y) - m_\theta(X)$.

In this paper, we are interested in the estimation of the probability density function (p.d.f.) f_ε of the residual term $\varepsilon = \Lambda_{\theta_o}(Y) - m(X)$. To this end, we first obtain the estimators $\widehat{\theta}$ and $\widehat{m}_\theta$ of the parameter θ_o and the function m_θ, and second, form the semiparametric regression residuals $\widehat{\varepsilon}_i(\widehat{\theta}) = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\theta}}(X_i)$. To estimate θ_o we use a profile likelihood (PL) approach, developed in Linton et al. (2008), whereas $\widehat{m}_\theta$ is estimated by means of a Nadaraya-Watson-type estimator (Nadaraya (1964), Watson (1964)). To our knowledge, the estimation of the density of ε in model (**??**) has not yet been investigated in the statistical literature. Estimating the error density in the semiparametric transformation model (SPT) $\Lambda_{\theta_o}(Y) = m(X) + \varepsilon$ may be very useful in various regression problems. First, taking transformations of the data may induce normality and error variance homogeneity in the transformed model. So the estimation of the error density in the transformed model may be used for testing these hypotheses; it may also be used for goodness-of-fit tests of a specified error distribution in a parametric or nonparametric regression setting. Some examples can be found in Akritas and Van Keilegom (2001), Cheng and Sun (2008), but with $\Lambda_{\theta_o} \equiv id$, i.e. the response is not transformed. Next, the estimation of the error density in the above model can be useful for testing the symmetry of the residual distribution. See Ahmad and Li (1997), Dette et al. (2002), Neumeyer and Dette (2007) and references therein, in the case $\Lambda_{\theta_o} \equiv id$. Under this model, Escanciano and Jacho-Chavez (2012) considered the estimation of the (marginal) density of the response Y via the estimation of the error density. Another application of the estimation of the error density in the SPT model is the forecasting of $\Lambda_{\theta_o}(Y)$ by

means of the mode approach, since the mode of the p.d.f. of $\Lambda_{\theta_o}(Y)$ given $X = x$ is $m(x) + \arg\max_{e\in\mathbb{R}} f_\varepsilon(e)$, where f_ε is the p.d.f. of the error term ε.

Taking transformations of the data has been an important part of statistical practice for many years. A major contribution to this methodology was made by Box and Cox (1964), who proposed a parametric power family of transformations that includes the logarithm and the identity. They suggested that the power transformation, when applied to the dependent variable in a linear regression model, might induce normality and homoscedasticity. Lots of effort has been devoted to the investigation of the Box-Cox transformation since its introduction. See, for example, Amemiya (1985), Horowitz (1998), Chen et al. (2002), Shin (2008), and Fitzenberger et al. (2010). Other dependent variable transformations have been suggested, for example, the Zellner and Revankar (1969) transform and the Bickel and Doksum (1981) transform. The transformation methodology has been quite successful and a large literature exists on this topic for parametric models. See Carroll and Ruppert (1988) and Sakia (1992) and references therein.

The estimation of (functionals of) the error distribution and density under simplified versions of model (**??**) has received considerable attention in the statistical literature in recent years. Akritas and Van Keilegom (2001) estimated the cumulative distribution function of the regression error in a heteroscedastic model with univariate covariates. The estimator they proposed is based on nonparametrically estimated regression residuals. The weak convergence of their estimator was proved. The results obtained by Akritas and Van Keilegom (2001) have been generalized by Neumeyer and Van Keilegom (2010) to the case of multivariate covariates. Müller et al. (2004) investigated linear functionals of the error distribution in nonparametric regression. Cheng (2005) established the asymptotic normality of an estimator of the error density based on estimated residuals. The estimator he proposed is constructed by splitting the sample into two parts: the first part is used for the estimation of the residuals, while the second part of the sample is used for the construction of the error density estimator. Efromovich (2005) proposed an adaptive estimator of the error density, based on a density estimator proposed by Pinsker (1980). Finally, Samb (2011) also considered the estimation of the error density, but his approach is more closely related to the one in Akritas and Van Keilegom (2001).

In order to achieve the objective of this paper, namely the estimation of the error density under model (**??**), we first need to estimate the transformation parameter θ_o. To this end, we make use of the results in Linton et al. (2008). In the latter paper, the authors first discuss the nonparametric identification of model (**??**), and second, estimate the transformation parameter θ_o under the considered model. For the estimation of this parameter, they propose two approaches. The first approach uses a semiparametric profile likelihood

(PL) estimator, while the second is based on a mean squared distance from independence-estimator (MD) using the estimated distributions of X, ε and (X, ε). Linton et al. (2008) derived the asymptotic distributions of their estimators under certain regularity conditions, and proved that both estimators of θ_o are root-n consistent. The authors also showed that, in practice, the performance of the PL method is better than that of the MD approach. For this reason, the PL method will be considered in this paper for the estimation of θ_o.

The rest of the paper is organized as follows. Section 2 presents our estimator of the error density and groups some notations and technical assumptions. Section 3 describes the asymptotic results of the paper. A simulation study is given in Section 4, while Section 5 is devoted to some general conclusions. Finally, the proofs of the asymptotic results are collected in Section 6.

2 Definitions and assumptions

2.1 Construction of the estimators

The approach proposed here for the estimation of f_ε is based on a two-steps procedure. In a first step, we estimate the finite dimensional parameter θ_o. This parameter is estimated by the profile likelihood (PL) method, developed in Linton et al. (2008). The basic idea of this method is to replace all unknown expressions in the likelihood function by their nonparametric kernel estimates. Under model (**??**), we have

$$\mathbb{P}\left(Y \leq y|X\right) = \mathbb{P}\left(\Lambda_{\theta_o}(Y) \leq \Lambda_{\theta_o}(y)|X\right) = \mathbb{P}\left(\varepsilon_{\theta_o} \leq \Lambda_{\theta_o}(y) - m_{\theta_o}(X)|X\right) = F_\varepsilon\left(\Lambda_{\theta_o}(y) - m_{\theta_o}(X)\right).$$

Here, $F_\varepsilon(t) = \mathbb{P}(\varepsilon \leq t)$, and so

$$f_{Y|X}(y|x) = f_\varepsilon\left(\Lambda_{\theta_o}(y) - m_{\theta_o}(x)\right)\Lambda'_{\theta_o}(y),$$

where f_ε and $f_{Y|X}$ are the densities of ε, and of Y given X, respectively. Then, the log likelihood function is

$$\sum_{i=1}^{n}\left\{\log f_{\varepsilon_\theta}(\Lambda_\theta(Y_i) - m_\theta(X_i)) + \log\Lambda'_\theta(Y_i)\right\}, \quad \theta \in \Theta,$$

where f_{ε_θ} is the density function of ε_θ. Now, let

$$\widehat{m}_\theta(x) = \frac{\sum_{j=1}^{n}\Lambda_\theta(Y_j)K_1\left(\frac{X_j - x}{h}\right)}{\sum_{j=1}^{n}K_1\left(\frac{X_j - x}{h}\right)} \tag{2}$$

be the Nadaraya-Watson estimator of $m_\theta(x)$, and let

$$\widehat{f}_{\varepsilon_\theta}(t) = \frac{1}{ng}\sum_{i=1}^{n}K_2\left(\frac{\widehat{\varepsilon}_i(\theta) - t}{g}\right). \tag{3}$$

where $\widehat{\varepsilon}_i(\theta) = \Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)$. Here, K_1 and K_2 are kernel functions and h and g are bandwidth sequences. Then, the PL estimator of θ_o is defined by

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \left[\log \widehat{f}_{\varepsilon_\theta}(\Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)) + \log \Lambda'_\theta(Y_i)\right]. \tag{4}$$

Recall that $\widehat{m}_\theta(X_i)$ converges to $m_\theta(X_i)$ at a slower rate for those X_i which are close to the boundary of the support $\mathcal{X}$ of the covariate X. That is why we assume implicitly that the proposed estimator (**??**) of θ_o trims the observations X_i outside a subset $\mathcal{X}_0$ of $\mathcal{X}$. Note that we keep the root-n consistency of $\widehat{\theta}$ proved in Linton et al. (2008) by trimming the covariates outside $\mathcal{X}_0$. But in this case, the resulting asymptotic variance is different to the one obtained in the latter paper.

In a second step, we use the above estimator $\widehat{\theta}$ to build the estimated residuals $\widehat{\varepsilon}_i(\widehat{\theta}) = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\theta}}(X_i)$. Then, our proposed estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$ of $f_\varepsilon(t)$ is defined by

$$\widehat{f}_{\widehat{\varepsilon}}(t) = \frac{1}{nb}\sum_{i=1}^{n} K_3\left(\frac{\widehat{\varepsilon}_i(\widehat{\theta}) - t}{b}\right), \tag{5}$$

where K_3 is a kernel function and b is a bandwidth sequence, not necessarily the same as the kernel K_2 and the bandwidth g used in (**??**). Observe that this estimator is a feasible estimator in the sense that it does not depend on any unknown quantity, as is desirable in practice. This contrasts with the unfeasible ideal kernel estimator

$$\widetilde{f}_{\varepsilon}(t) = \frac{1}{nb}\sum_{i=1}^{n} K_3\left(\frac{\varepsilon_i - t}{b}\right), \tag{6}$$

which depends in particular on the unknown regression errors $\varepsilon_i = \varepsilon_i(\theta_o) = \Lambda_{\theta_o}(Y_i) - m(X_i)$. It is however intuitively clear that $\widehat{f}_{\widehat{\varepsilon}}(t)$ and $\widetilde{f}_{\varepsilon}(t)$ will be very close for n large enough, as will be illustrated by the results given in Section 3.

2.2 Notations

When there is no ambiguity, we use ε and m to indicate ε_{θ_o} and m_{θ_o}. Moreover, $\mathcal{N}(\theta_o)$ represents a neighborhood of θ_o. For the kernel K_j $(j = 1, 2, 3)$, let $\mu(K_j) = \int v^2 K_j(v)dv$ and let $K_j^{(p)}$ be the pth derivative of K_j. For any function $\varphi_\theta(y)$, denote $\dot{\varphi}_\theta(y) = \partial\varphi_\theta(y)/\partial\theta = (\partial\varphi_\theta(y)/\partial\theta_1, \ldots, \partial\varphi_\theta(y)/\partial\theta_p)^t$ and $\varphi'_\theta(y) = \partial\varphi_\theta(y)/\partial y$. Also, let $\|A\| = (A^tA)^{1/2}$ be the Euclidean norm of any vector A.

For any functions $\widetilde{m}$, r, f, φ and q, and any $\theta \in \Theta$, let $s = (\widetilde{m}, r, f, \varphi, q)$, $s_\theta = (m_\theta, \dot{m}_\theta, f_{\varepsilon_\theta}, f'_{\varepsilon_\theta}, \dot{f}_{\varepsilon_\theta})$, $\varepsilon_i(\theta, \widetilde{m}) = \Lambda_\theta(Y_i) - \widetilde{m}(X_i)$, and define

$$G_n(\theta, s) = n^{-1}\sum_{i=1}^{n}\left\{\frac{1}{f\{\varepsilon_i(\theta,\widetilde{m})\}}\left[\varphi\{\varepsilon_i(\theta,\widetilde{m})\}\{\dot{\Lambda}_\theta(Y_i) - r(X_i)\} + q\{\varepsilon_i(\theta,\widetilde{m})\}\right] + \frac{\dot{\Lambda}'_\theta(Y_i)}{\Lambda'_\theta(Y_i)}\right\},$$

$G(\theta, s) = \mathbb{E}[G_n(\theta, s)]$ and $\mathcal{G}(\theta_o, s_{\theta_o}) = \frac{\partial}{\partial\theta} G(\theta, s_\theta)\downarrow_{\theta=\theta_o}$.

2.3 Technical assumptions

The assumptions we need for the asymptotic results are listed below for convenient reference.

(**A1**) The function K_j $(j = 1, 2, 3)$ is symmetric, has compact support, $\int v^k K_j(v)dv = 0$ for $k = 1, \ldots, q_j - 1$ and $\int v^{q_j} K_j(v)dv \neq 0$ for some $q_j \geq 4$, K_j is twice continuously differentiable, and $\int K_3^{(1)}(v)dv = 0$.

(**A2**) The bandwidth sequences h, g and b satisfy $nh^{2q_1} = o(1)$, $ng^{2q_2} = o(1)$ (where q_1 and q_2 are defined in (A1)), $(nb^5)^{-1} = O(1)$, $nb^3h^2(\log h^{-1})^{-2} \to \infty$ and $ng^6(\log g^{-1})^{-2} \to \infty$.

(**A3**) (i) The support $\mathcal{X}$ of the covariate X is a compact subset of $\mathbb{R}$, and $\mathcal{X}_0$ is a subset with non empty interior, whose closure is in the interior of $\mathcal{X}$.
(ii) The density f_X is bounded away from zero and infinity on $\mathcal{X}$, and has continuous second order partial derivatives on $\mathcal{X}$.

(**A4**) The function $m_\theta(x)$ is twice continuously differentiable with respect to θ on $\mathcal{X} \times \mathcal{N}(\theta_0)$, and the functions $m_\theta(x)$ and $\dot{m}_\theta(x)$ are q_1 times continuously differentiable with respect to x on $\mathcal{X} \times \mathcal{N}(\theta_0)$. All these derivatives are bounded, uniformly in $(x, \theta) \in \mathcal{X} \times \mathcal{N}(\theta_o)$.

(**A5**) The error $\varepsilon = \Lambda_{\theta_o}(Y) - m(X)$ has finite fourth moment and is independent of X.

(**A6**) The distribution $F_{\varepsilon_\theta}(t)$ is $q_3 + 1$ (respectively three) times continuously differentiable with respect to t (respectively θ), and

$$\sup_{\theta,t} \left| \frac{\partial^{k+\ell}}{\partial t^k \partial\theta_1^{\ell_1} \ldots \partial\theta_p^{\ell_p}} F_{\varepsilon_\theta}(t) \right| < \infty$$

for all k and ℓ such that $0 \leq k + \ell \leq 2$, where $\ell = \ell_1 + \ldots + \ell_p$ and $\theta = (\theta_1, \ldots, \theta_p)^t$.

(**A7**) The transformation $\Lambda_\theta(y)$ is three times continuously differentiable with respect to both θ and y, and there exists a $\alpha > 0$ such that

$$\mathbb{E}\left[\sup_{\theta' : \|\theta' - \theta\| \leq \alpha} \left| \frac{\partial^{k+\ell}}{\partial y^k \partial\theta_1^{\ell_1} \ldots \partial\theta_p^{\ell_p}} \Lambda_{\theta'}(Y) \right| \right] < \infty$$

for all $\theta \in \Theta$, and for all k and ℓ such that $0 \leq k + \ell \leq 3$, where $\ell = \ell_1 + \ldots + \ell_p$ and $\theta = (\theta_1, \ldots, \theta_p)^t$. Moreover, $\sup_{x \in \mathcal{X}} \mathbb{E}[\dot{\Lambda}_{\theta_o}^4(Y)|X = x] < \infty$.

(**A8**) For all $\eta > 0$, there exists $\epsilon(\eta) > 0$ such that

$$\inf_{\|\theta - \theta_o\| > \eta} \|G(\theta, s_\theta)\| \geq \epsilon(\eta) > 0.$$

Moreover, the matrix $\mathcal{G}(\theta_o, s_{\theta_o})$ is non-singular.

(**A9**) (i) $\mathbb{E}(\Lambda_{\theta_o}(Y)) = 1$, $\Lambda_{\theta_o}(0) = 0$ and the set $\{x \in \mathcal{X}_0 : m'(x) \neq 0\}$ has nonempty interior.

(ii) Assume that $\phi(x,t) = \dot{\Lambda}_{\theta_o}(\Lambda_{\theta_o}^{-1}(m(x)+t))f_\varepsilon(t)$ is continuously differentiable with respect to t for all x and that

$$\sup_{s:|t-s|\leq\delta} \mathbb{E}\left|\frac{\partial\phi}{\partial s}(X,s)\right| < \infty. \tag{7}$$

for all $t \in \mathbb{R}$ and for some $\delta > 0$.

Assumptions (A1), part of (A2), (A3)(ii), (A4) and (A6), (A7) and (A8) are used by Linton et al. (2008) to show that the PL estimator $\widehat{\theta}$ of θ_o is root n-consistent. The differentiability of K_j up to second order imposed in assumption (A1) is used to expand the two-steps kernel estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$ in (**??**) around the unfeasible one $\widetilde{f}_{\varepsilon}(t)$. Assumptions (A3)(ii) and (A4) impose that all the functions to be estimated have bounded derivatives. The last assumption in (A2) is useful for obtaining the uniform convergence of the Nadaraya-Watson estimator of m_{θ_o} in (**??**) (see for instance Einmahl and Mason (2005)). This assumption is also needed in the study of the difference between the feasible estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$ and the unfeasible estimator $\widetilde{f}_{\varepsilon}(t)$. Finally, (A9)(i) is needed for identifying the model (see Vanhems and Van Keilegom (2011)).

3 Asymptotic results

In this section we are interested in the asymptotic behavior of the estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$. To this end, we first investigate its asymptotic representation, which will be needed to show its asymptotic normality.

Theorem 1. *Assume* (A1)-(A9). *Then,*

$$\widehat{f}_{\widehat{\varepsilon}}(t) - f_\varepsilon(t) = \frac{1}{nb}\sum_{i=1}^{n} K_3\left(\frac{\varepsilon_i - t}{b}\right) - f_\varepsilon(t) + R_n(t),$$

where $R_n(t) = o_{\mathbb{P}}\left((nb)^{-1/2}\right)$ *for all* $t \in \mathbb{R}$.

This result is important, since it shows that, provided the bias term is negligible, the estimation of θ_o and $m(\cdot)$ has asymptotically no effect on the behavior of the estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$. Therefore, this estimator is asymptotically equivalent to the unfeasible estimator $\widetilde{f}_{\varepsilon}(t)$, based on the unknown true errors $\varepsilon_1, \ldots, \varepsilon_n$.

Our next result gives the asymptotic normality of the estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$.

Theorem 2. *Assume* (A1)-(A9). *In addition, assume that* $nb^{2q_3+1} = O(1)$. *Then,*

$$\sqrt{nb}\left(\widehat{f}_{\widehat{\varepsilon}}(t) - \overline{f}_{\varepsilon}(t)\right) \xrightarrow{d} N\left(0, f_\varepsilon(t)\int K_3^2(v)dv\right),$$

where

$$\overline{f}_{\varepsilon}(t) = f_{\varepsilon}(t) + \frac{b^{q_3}}{q_3!} f_{\varepsilon}^{(q_3)}(t) \int v^{q_3} K_3(v) dv.$$

The proofs of Theorems **??** and **??** are given in Section **??**.

4 Simulations

In this section, we investigate the performance of our method for different models and different sample sizes. Consider

$$\Lambda_{\theta_o}(Y) = b_0 + b_1 X^2 + b_2 \sin(\pi X) + \sigma_e \varepsilon, \tag{8}$$

where Λ_θ is the Box-Cox (1964) transformation

$$\Lambda_\theta(y) = \begin{cases} \frac{y^\theta - 1}{\theta}, & \theta \neq 0, \\ \log(y), & \theta = 0, \end{cases}$$

X is uniformly distributed on the interval $[-1, 1]$, and ε is independent of X. We carry out simulations for two cases : in the first case, ε has a standard normal distribution and, in the second case, the distribution of ε is the mixture of the normal distributions $N(-1.5, 0.25)$ and $N(1.5, 0.25)$ with equal weights. To make computations easier, error distributions are truncated at -3 and 3 (i.e., put to 0 outside the interval $[-3, 3]$). We study three different model settings. For each of them, $b_2 = b_0 - 3\sigma_e$. The other parameters are chosen as follows:

Model 1:	$b_0 = 6.5$,	$b_1 = 5$,	$\sigma_e = 1.5$;
Model 2:	$b_0 = 4.5$,	$b_1 = 3.5$,	$\sigma_e = 1$;
Model 3:	$b_0 = 2.5$,	$b_1 = 2.5$,	$\sigma_e = 0.5$.

Our simulations are performed for $\theta_0 = 0$, 0.5 and 1. We use the Epanechnikov kernel $K(x) = \frac{3}{4}\left(1 - x^2\right) \mathbb{1}\left(|x| \leq 1\right)$ for both the estimator of the regression function and the density function. The results are based on 100 random samples of size $n = 100$ and $n = 200$. For the estimation of θ_0 and $f_\varepsilon(t)$, we proceed as follows. Let

$$L_\theta(h, g) = \sum_{i=1}^{n} \left[\log \widehat{f}_{\varepsilon_\theta}(\widehat{\varepsilon}_i(\theta, h)) + \log \Lambda'_\theta(Y_i)\right],$$

where $\widehat{\varepsilon}_i(\theta, h) = \Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i, h)$ and $\widehat{m}_\theta(x, h)$ denotes $\widehat{m}_\theta(x)$ constructed with bandwidth h. This function will be maximized with respect to θ for given (optimal) values of (h, g). For each value of θ, $h^*(\theta)$ is obtained

by least squares cross-validation,

$$h^*(\theta) = \arg\max_h \sum_{i=1}^{n} (\Lambda_\theta(Y_i) - \widehat{m}_{-i,\theta}(X_i))^2,$$

where

$$\widehat{m}_{-i,\theta}(X_i) = \frac{\sum_{j=1,j\neq i}^{n} \Lambda_\theta(Y_j) K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j=1,j\neq i}^{n} K\left(\frac{X_j - X_i}{h}\right)}$$

and g can be chosen with a classical bandwidth selection rule for kernel density estimation. Here, for simplicity, the normal rule is used ($\widehat{g}(\theta) = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$, where $\widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$ is the classical empirical estimator of the standard deviation based on $\widehat{\varepsilon}_i(\theta, h^*(\theta))$, $i = 1, \ldots, n$). The solution

$$\widehat{\theta} = \arg\max_\theta L_\theta(h^*(\theta), \widehat{g}(\theta))$$

is therefore obtained iteratively (maximization problems are solved with the function 'optimize' in R with $h \in [0, 2]$ and $\theta \in [-20, 20]$) and the estimator of $f_\varepsilon(t)$ is finally given by

$$\widehat{f}_{\widehat{\varepsilon}}(t) \quad = \quad \frac{1}{n\widehat{g}(\widehat{\theta})} \sum_{i=1}^{n} K\left(\frac{\widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta})) - t}{\widehat{g}(\widehat{\theta})}\right).$$

Tables **??**, **??** and **??** show the mean squared error (MSE) of the estimator $\widehat{f}_{\widehat{\varepsilon}}(t)$ of the standardized (pseudo-estimated) error $\tilde{\varepsilon} = \left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X)\right)/\sigma_e$ (with known σ_e), for $t = -1$, 0 and 1 (respectively $t = -1.5$, -1, 0, 1 and 1.5) and for the unimodal (respectively bimodal) normal error distribution. Tables **??** and **??** show the integrated mean squared error (IMSE) of the estimator $\widehat{f}_{\widehat{\varepsilon}}(\cdot)$ for both error distributions, where the integration is done over the interval $[-3, 3]$. As expected, in both cases, estimation is better for the normal density than for the mixture of two normals, and estimation improves when n increases, and in most cases, when σ_e decreases. In particular, this can be observed from Tables **??** and **??**. The limiting case $\theta_0 = 0$ (the logarithmic transformation) seems to be more easily captured, especially when the error is normally distributed. In general, we observe from Tables **??**, **??**, **??** that estimation is poorer near local maxima and minima of the density, which is not uncommon for kernel smoothing methods. This also suggests that the choice of the smoothing parameters is important and should be the object of further investigation.

Model	θ_0		$n = 100$			$n = 200$		
			$\widehat{f_{\tilde{\epsilon}}}(-1)$	$\widehat{f_{\tilde{\epsilon}}}(0)$	$\widehat{f_{\tilde{\epsilon}}}(1)$	$\widehat{f_{\tilde{\epsilon}}}(-1)$	$\widehat{f_{\tilde{\epsilon}}}(0)$	$\widehat{f_{\tilde{\epsilon}}}(1)$
		Bias	-.0421	-.0206	-.0123	-.0183	.0196	.0004
	$\theta_0 = 0$	Var	.0006	.0206	.0017	.0008	.0116	.0008
		MSE	.0024	.0211	.0018	.0011	.0120	.0008
$b_0 = 6.5$		Bias	-.0621	.0469	-.0631	-.0521	.0309	-.0262
$b_1 = 5$	$\theta_0 = 0.5$	Var	.0051	.1624	.0061	.0030	.1555	.0066
$\sigma_0 = 1.5$		MSE	.0089	.1646	.0101	.0057	.1565	.0073
		Bias	-.0874	.0806	-.0885	-.0530	.1063	-.0737
	$\theta_0 = 1$	Var	.0073	.2261	.0089	.0049	.1152	.0032
		MSE	.0149	.2326	.0168	.0077	.1265	.0086
		Bias	-.0029	-.0953	-.0232	-.0419	.0627	-.0118
	$\theta_0 = 0$	Var	.0019	.0142	.0023	.0004	.0130	.0010
		MSE	.0019	.0233	.0028	.0022	.0169	.0012
$b_0 = 4.5$		Bias	-.0522	.0476	-.0435	-.0228	-.0193	-.0146
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	.0041	.1184	.0062	.0017	.0333	.0020
$\sigma_0 = 1$		MSE	.0068	.1207	.0081	.0022	.0337	.0022
		Bias	-.0703	.1816	-.0837	-.0425	.0240	-.0413
	$\theta_0 = 1$	Var	.0049	.2497	.0045	.0023	.0519	.0028
		MSE	.0098	.2827	.0114	.0041	.0525	.0045
		Bias	-.0323	-.0053	-.0008	-.0073	.0306	-.0373
	$\theta_0 = 0$	Var	.0006	.0148	.0011	.0005	.0063	.0002
		MSE	.0017	.0148	.0011	.0005	.0072	.0016
$b_0 = 2.5$		Bias	-.0304	.0156	-.0289	-.0214	.0223	-.0164
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	.0014	.0266	.0020	.0008	.0129	.0008
$\sigma_0 = 0.5$		MSE	.0024	.0268	.0028	.0012	.0134	.0011
		Bias	-.0252	.0411	-.0308	-.0442	.0836	-.0303
	$\theta_0 = 1$	Var	.0020	.0415	.0042	.0007	.0256	.0014
		MSE	.0026	.0432	.0052	.0026	.0325	.0023

Table 1: $MSE(\widehat{f_{\tilde{\varepsilon}}}(t))$ for different models, values of t and sample sizes, when $f_{\varepsilon}(\cdot)$ is a standard normal density.

Model	θ_0	$n = 100$	$n = 200$
$b_0 = 6.5$	$\theta_0 = 0$	.0042	.0023
$b_1 = 5$	$\theta_0 = 0.5$	.0161	.0106
$\sigma_0 = 1.5$	$\theta_0 = 1$	.0237	.0129
$b_0 = 4.5$	$\theta_0 = 0$	.0060	.0029
$b_1 = 3.5$	$\theta_0 = 0.5$	.0125	.0053
$\sigma_0 = 1$	$\theta_0 = 1$	.0191	.0075
$b_0 = 2.5$	$\theta_0 = 0$	.0027	.0015
$b_1 = 2.5$	$\theta_0 = 0.5$	.0048	.0026
$\sigma_0 = 0.5$	$\theta_0 = 1$	.0114	.0036

Table 2: $IMSE(\widehat{f}_{\widehat{\varepsilon}})$ for different models and sample sizes, when $f_\varepsilon(\cdot)$ is a standard normal density.

5 Conclusions

In this paper we have studied the estimation of the density of the error in a semiparametric transformation model. The regression function in this model is unspecified (except for some smoothness assumptions), whereas the transformation (of the dependent variable in the model) is supposed to belong to a parametric family of monotone transformations. The proposed estimator is a kernel-type estimator, and we have shown its asymptotic normality. The finite sample performance of the estimator is illustrated by means of a simulation study.

It would be interesting to explore various possible applications of the results in this paper. For example, one could use the results on the estimation of the error density to test hypotheses concerning e.g. the normality of the errors, the homoscedasticity of the error variance, or the linearity of the regression function, all of which are important features in the context of transformation models.

6 Proofs

Proof of Theorem ??. Write

$$\widehat{f}_{\widehat{\varepsilon}}(t) - f_\varepsilon(t) = [\widehat{f}_\varepsilon(t) - f_\varepsilon(t)] + [\widehat{f}_{\widehat{\varepsilon}}(t) - \widehat{f}_\varepsilon(t)],$$

where

$$\widehat{f}_\varepsilon(t) = \frac{1}{nb} \sum_{i=1}^{n} K_3 \left(\frac{\widehat{\varepsilon}_i - t}{b} \right)$$

Model	θ_0		$n = 100$				
			$\widehat{f}_{\tilde{\varepsilon}}(-1.5)$	$\widehat{f}_{\tilde{\varepsilon}}(-1)$	$\widehat{f}_{\tilde{\varepsilon}}(0)$	$\widehat{f}_{\tilde{\varepsilon}}(1)$	$\widehat{f}_{\tilde{\varepsilon}}(1.5)$
		Bias	-.1955	-.0292	.1671	-.0359	-.2069
	$\theta_0 = 0$	Var	.0003	.0010	.0013	.0012	.0005
		MSE	.0386	.0018	.0293	.0024	.0433
$b_0 = 6.5$		Bias	-.1854	-.0004	.1252	-.0086	-.1913
$b_1 = 5$	$\theta_0 = 0.5$	Var	.0021	.0053	.0017	.0059	.0021
$\sigma_0 = 1.5$		MSE	.0365	.0053	.0174	.0060	.0387
		Bias	-.2055	-.0046	.1641	-.0188	-.2173
	$\theta_0 = 1$	Var	.0033	.0065	.0167	.0061	.0027
		MSE	.0455	.0065	.0436	.0065	.0499
		Bias	-.1665	.0514	.1921	-.0875	-.2354
	$\theta_0 = 0$	Var	.0004	.0014	.0010	.0008	.0005
		MSE	.0282	.0040	.0379	.0084	.0589
$b_0 = 4.5$		Bias	-.1973	-.0235	.1584	-.0066	-.1892
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	.0007	.0026	.0016	.0038	.0012
$\sigma_0 = 1$		MSE	.0396	.0031	.0267	.0038	.0370
		Bias	-.2025	-.0271	.1659	.0221	-.1902
	$\theta_0 = 1$	Var	.0015	.0039	.0044	.0039	.0017
		MSE	.0425	.0046	.0319	.0044	.0379
		Bias	-.1544	.0698	.1915	-.1296	-.2547
	$\theta_0 = 0$	Var	.0003	.0009	.0006	.0004	.0007
		MSE	.0242	.0057	.0372	.0172	.0656
$b_0 = 2.5$		Bias	-.1924	-.0501	.1341	.0317	-.1459
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	.0004	.0011	.0007	.0021	.0005
$\sigma_0 = 0.5$		MSE	.0374	.0036	.0187	.0031	.0218
		Bias	-.1654	.0123	.1289	-.0642	-.1944
	$\theta_0 = 1$	Var	.0005	.0017	.0010	.0022	.0013
		MSE	.0279	.0019	.0167	.0063	.0391

Table 3: $MSE(\widehat{f}_{\tilde{\varepsilon}}(t))$ for different models, values of t and $n = 100$, when $f_{\varepsilon}(\cdot)$ is a mixture of two normal densities ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

Model	θ_0		$n = 200$				
			$\widehat{f}_{\tilde{\varepsilon}}(-1.5)$	$\widehat{f}_{\tilde{\varepsilon}}(-1)$	$\widehat{f}_{\tilde{\varepsilon}}(0)$	$\widehat{f}_{\tilde{\varepsilon}}(1)$	$\widehat{f}_{\tilde{\varepsilon}}(1.5)$
		Bias	-.1578	-.0132	.1103	-.0212	-.1665
	$\theta_0 = 0$	Var	.0003	.0009	.0002	.0010	.0003
		MSE	.0252	.0011	.0123	.0015	.0281
$b_0 = 6.5$		Bias	-.1425	.0372	.0960	-.0193	-.1652
$b_1 = 5$	$\theta_0 = 0.5$	Var	.0009	.0038	.0005	.0039	.0019
$\sigma_0 = 1.5$		MSE	.0212	.0052	.0097	.0043	.0285
		Bias	-.1697	-.0077	.1019	-.0213	-.1769
	$\theta_0 = 1$	Var	.0014	.0047	.0007	.0051	.0018
		MSE	.0302	.0048	.0111	.0056	.0331
		Bias	-.1511	-.0022	.0980	-.0348	-.1681
	$\theta_0 = 0$	Var	.0002	.0007	.0001	.0009	.0004
		MSE	.0230	.0007	.0098	.0021	.0286
$b_0 = 4.5$		Bias	-.1712	-.0287	.1092	.0099	-.1538
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	.0005	.0019	.0004	.0025	.0005
$\sigma_0 = 1$		MSE	.0298	.0028	.0123	.0026	.0242
		Bias	-.1278	.0323	.0630	-.0228	-.1532
	$\theta_0 = 1$	Var	.0009	.0038	.0002	.0038	.0015
		MSE	.0173	.0048	.0042	.0043	.0250
		Bias	-.1430	.0008	.0915	-.0581	-.1749
	$\theta_0 = 0$	Var	.0001	.0004	.0001	.0005	.0004
		MSE	.0205	.0004	.0085	.0039	.0310
$b_0 = 2.5$		Bias	-.1406	.0245	.1067	-.0485	-.1673
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	.0001	.0008	.0002	.0012	.0006
$\sigma_0 = 0.5$		MSE	.0199	.0014	.0116	.0035	.0286
		Bias	-.1551	-.0291	.0839	.0013	-.1436
	$\theta_0 = 1$	Var	.0003	.0010	.0001	.0013	.0003
		MSE	.0244	.0019	.0072	.0013	.0210

Table 4: $MSE(\widehat{f}_{\tilde{\varepsilon}}(t))$ for different models, values of t and $n = 200$, when $f_{\varepsilon}(\cdot)$ is a mixture of two normal densities ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

Model	θ_0	$n = 100$	$n = 200$
$b_0 = 6.5$	$\theta_0 = 0$	.0148	.0089
$b_1 = 5$	$\theta_0 = 0.5$	.0158	.0106
$\sigma_0 = 1.5$	$\theta_0 = 1$	.0219	.0119
$b_0 = 4.5$	$\theta_0 = 0$	.0184	.0082
$b_1 = 3.5$	$\theta_0 = 0.5$	.0157	.0099
$\sigma_0 = 1$	$\theta_0 = 1$	.0186	.0083
$b_0 = 2.5$	$\theta_0 = 0$	.0199	.0079
$b_1 = 2.5$	$\theta_0 = 0.5$	.0118	.0087
$\sigma_0 = 0.5$	$\theta_0 = 1$	.0123	.0078

Table 5: $IMSE(\widehat{f}_{\widehat{\varepsilon}})$ for different models and sample sizes, when $f_\varepsilon(\cdot)$ is a mixture of two normal densities ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

and $\widehat{\varepsilon}_i = \Lambda_{\theta_o}(Y_i) - \widehat{m}_{\theta_o}(X_i)$, $i = 1, \ldots, n$. In a completely similar way as was done for Lemma A.1 in Linton et al. (2008), it can be shown that

$$\widehat{f}_\varepsilon(t) - f_\varepsilon(t) = \frac{1}{nb} \sum_{i=1}^{n} K_3 \left(\frac{\varepsilon_i - t}{b} \right) - f_\varepsilon(t) + o_{\mathbb{P}}((nb)^{-1/2}) \tag{9}$$

for all $t \in \mathbb{R}$. Note that the remainder term in Lemma A.1 in the above paper equals a sum of i.i.d. terms of mean zero, plus a $o_{\mathbb{P}}(n^{-1/2})$ term. Hence, the remainder term in that paper is $O_{\mathbb{P}}(n^{-1/2})$, whereas we write $o_{\mathbb{P}}((nb)^{-1/2})$ in (**??**). Therefore, the result of the theorem follows if we prove that $\widehat{f}_{\widehat{\varepsilon}}(t) - \widehat{f}_\varepsilon(t) = o_{\mathbb{P}}((nb)^{-1/2})$. To this end, write

$$\begin{aligned}
&\widehat{f}_{\widehat{\varepsilon}}(t) - \widehat{f}_\varepsilon(t) \\
&= \frac{1}{nb^2} \sum_{i=1}^{n} (\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o)) K_3^{(1)} \left(\frac{\widehat{\varepsilon}_i(\theta_o) - t}{b} \right) \\
&\quad + \frac{1}{2nb^3} \sum_{i=1}^{n} (\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o))^2 K_3^{(2)} \left(\frac{\widehat{\varepsilon}_i(\theta_o) + \beta(\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o)) - t}{b} \right),
\end{aligned}$$

for some $\beta \in (0, 1)$. In what follows, we will show that each of the terms above is $o_{\mathbb{P}}((nb)^{-1/2})$. First consider the last term of (**??**). Since $\Lambda_\theta(y)$ and $\widehat{m}_\theta(x)$ are both twice continuously differentiable with respect to θ, the second order Taylor expansion gives, for some θ_1 between θ_o and $\widehat{\theta}$ (to simplify the notations, we assume

here that $p = \dim(\theta) = 1$),

$$\begin{aligned}
&\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o) \\
&\quad = \Lambda_{\widehat{\theta}}(Y_i) - \Lambda_{\theta_o}(Y_i) - \big(\widehat{m}_{\widehat{\theta}}(X_i) - \widehat{m}_{\theta_o}(X_i)\big) \\
&\quad = (\widehat{\theta} - \theta_o)(\dot{\Lambda}_{\theta_o}(Y_i) - \dot{\widehat{m}}_{\theta_o}(X_i)) + \frac{1}{2}(\widehat{\theta} - \theta_o)^2(\ddot{\Lambda}_{\theta_1}(Y_i) - \ddot{\widehat{m}}_{\theta_1}(X_i)).
\end{aligned}$$

Therefore, since $\widehat{\theta} - \theta_o = o_{\mathbb{P}}((nb)^{-1/2})$ by Theorem 4.1 in Linton et al. (2008) (as before, we work with a slower rate than what is shown in the latter paper, since this leads to weaker conditions on the bandwidths), assumptions (A2) and (A7) imply that

$$\frac{1}{nb^3}\sum_{i=1}^{n}(\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o))^2 K_3^{(2)}\left(\frac{\widehat{\varepsilon}_i(\theta_o) + \beta(\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o)) - t}{b}\right) = O_{\mathbb{P}}\left((nb^3)^{-1}\right),$$

which is $o_{\mathbb{P}}((nb)^{-1/2})$, since $(nb^5)^{-1} = O(1)$ under (A2). For the first term of (**??**), the decomposition of $\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o)$ given above yields

$$\begin{aligned}
&\frac{1}{nb^2}\sum_{i=1}^{n}(\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o))K_3^{(1)}\left(\frac{\widehat{\varepsilon}_i(\theta_o) - t}{b}\right) \\
&\quad = \frac{(\widehat{\theta} - \theta_o)}{nb^2}\sum_{i=1}^{n}(\dot{\Lambda}_{\theta_o}(Y_i) - \dot{\widehat{m}}_{\theta_o}(X_i))K_3^{(1)}\left(\frac{\widehat{\varepsilon}_i(\theta_o) - t}{b}\right) + o_{\mathbb{P}}((nb)^{-1/2}) \\
&\quad = \frac{(\widehat{\theta} - \theta_o)}{nb^2}\sum_{i=1}^{n}(\dot{\Lambda}_{\theta_o}(Y_i) - \dot{m}_{\theta_o}(X_i))K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right) + o_{\mathbb{P}}((nb)^{-1/2}), \qquad (10)
\end{aligned}$$

where the last equality follows from a Taylor expansion applied to $K_3^{(1)}$, the fact that

$$\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x) = O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}),$$

uniformly in $x \in \mathcal{X}_0$ by Lemma **??**, and the fact that $nhb^3(\log h^{-1})^{-1} \to \infty$ under (A2). Further, write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n}(\dot{\Lambda}_{\theta_o}(Y_i) - \dot{m}_{\theta_o}(X_i))K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right)\right] \\
&\quad = \sum_{i=1}^{n}\mathbb{E}\left[\dot{\Lambda}_{\theta_o}(Y_i)K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right)\right] - \sum_{i=1}^{n}\mathbb{E}\left[\dot{m}_{\theta_o}(X_i)\right]\mathbb{E}\left[K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right)\right] \\
&\quad = A_n - B_n.
\end{aligned}$$

We will only show that the first term above is $O(nb^2)$ for any $t \in \mathbb{R}$. The proof for the other term is similar. Let $\varphi(x,t) = \dot{\Lambda}_{\theta_o}(\Lambda_{\theta_o}^{-1}(m(x)+t))$ and set $\phi(x,t) = \varphi(x,t)f_\varepsilon(t)$. Then, applying a Taylor expansion to $\phi(x,\cdot)$,

it follows that (for some $\beta \in (0,1)$)

$$\begin{aligned} A_n &= \sum_{i=1}^{n} \mathbb{E}\left[\dot{\Lambda}_{\theta_o}\left(\Lambda_{\theta_o}^{-1}(m(X_i)+\varepsilon_i)\right) K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right)\right] \\ &= n\int\int \phi(x,e) K_3^{(1)}\left(\frac{e-t}{b}\right) f_X(x)dxde \\ &= nb\int\int \phi(x,t+bv) K_3^{(1)}(v) f_X(x)dxdv \\ &= nb\int\int \left[\phi(x,t) + bv\frac{\partial \phi}{\partial t}(x,t+\beta bv)\right] K_3^{(1)}(v) f_X(x)dxdv \\ &= nb^2\int\int v\frac{\partial \phi}{\partial t}(x,t+\beta bv) K_3^{(1)}(v) f_X(x)dxdv, \end{aligned}$$

since $\int K_3^{(1)}(v)dv = 0$, and this is bounded by $Knb^2 \sup_{s:|t-s|\le\delta} \mathbb{E}|\frac{\partial \phi}{\partial s}(X,s)| = O(nb^2)$ by assumption (A9)(ii). Hence, Tchebychev's inequality ensures that

$$\begin{aligned} &\frac{(\widehat{\theta}-\theta_o)}{b^2}\sum_{i=1}^{n}(\dot{\Lambda}_{\theta_o}(Y_i) - \dot{m}_{\theta_o}(X_i)) K_3^{(1)}\left(\frac{\varepsilon_i - t}{b}\right) \\ &= \frac{(\widehat{\theta}-\theta_o)}{nb^2} O_{\mathbb{P}}(nb^2 + (nb)^{1/2}) = o_{\mathbb{P}}((nb)^{-1/2}), \end{aligned}$$

since $nb^{3/2} \to \infty$ by (A2). Substituting this in (**??**), yields

$$\frac{1}{nb^2}\sum_{i=1}^{n}(\widehat{\varepsilon}_i(\widehat{\theta}) - \widehat{\varepsilon}_i(\theta_o)) K_3^{(1)}\left(\frac{\widehat{\varepsilon}_i(\theta_o) - t}{b}\right) = o_{\mathbb{P}}((nb)^{-1/2}),$$

for any $t \in \mathbb{R}$. This completes the proof. □

Proof of Theorem ??. It follows from Theorem **??** that

$$\widehat{f}_{\widehat{\varepsilon}}(t) - f_\varepsilon(t) = [\widetilde{f}_\varepsilon(t) - \mathbb{E}\widetilde{f}_\varepsilon(t)] + [\mathbb{E}\widetilde{f}_\varepsilon(t) - f_\varepsilon(t)] + o_{\mathbb{P}}((nb)^{-1/2}). \tag{11}$$

The first term on the right hand side of (**??**) is treated by Lyapounov's Central Limit Theorem (LCT) for triangular arrays (see e.g. Billingsley (1968), Theorem 7.3). To this end, let

$$\widetilde{f}_{in}(t) = \frac{1}{b}K_3\left(\frac{\varepsilon_i - t}{b}\right).$$

Then, under (A1), (A2) and (A5) it can be easily shown that

$$\frac{\sum_{i=1}^{n}\mathbb{E}\left|\widetilde{f}_{in}(t) - \mathbb{E}\widetilde{f}_{in}(t)\right|^3}{\left(\sum_{i=1}^{n}\mathrm{Var}\widetilde{f}_{in}(t)\right)^{3/2}} \le \frac{Cnb^{-2}f_\varepsilon(t)\int|K_3(v)|^3 dv + o\left(nb^{-2}\right)}{\left(nb^{-1}f_\varepsilon(t)\int K_3^2(v)dv + o\left(nb^{-1}\right)\right)^{3/2}} = O((nb)^{-1/2}) = o(1),$$

for some $C > 0$. Hence, the LCT ensures that

$$\frac{\widetilde{f}_\varepsilon(t) - \mathbb{E}\widetilde{f}_\varepsilon(t)}{\sqrt{\mathrm{Var}\widetilde{f}_\varepsilon(t)}} = \frac{\widetilde{f}_\varepsilon(t) - \mathbb{E}\widetilde{f}_\varepsilon(t)}{\sqrt{\frac{\mathrm{Var}\widetilde{f}_{1n}(t)}{n}}} \xrightarrow{d} N(0,1).$$

This gives

$$\sqrt{nb}\left(\widetilde{f}_\varepsilon(t) - \mathbb{E}\widetilde{f}_\varepsilon(t)\right) \xrightarrow{d} N\left(0, f_\varepsilon(t)\int K_3^2(v)dv\right). \tag{12}$$

For the second term of (**??**), straightforward calculations show that

$$\mathbb{E}\widetilde{f}_\varepsilon(t) - f_\varepsilon(t) = \frac{b^{q_3}}{q_3!} f_\varepsilon^{(q_3)}(t)\int v^{q_3}K_3(v)dv + o(b^{q_3}).$$

Combining this with (**??**) and (**??**), we obtain the desired result. □

Lemma 1. *Assume* (A1)-(A5) *and* (A7). *Then,*

$$\begin{aligned}
\sup_{x\in\mathcal{X}_0}|\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)| &= O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}),\\
\sup_{x\in\mathcal{X}_0}|\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)| &= O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}).
\end{aligned}$$

Proof. We will only show the proof for $\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)$, the proof for $\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)$ being very similar. Let $c_n = (nh)^{-1/2}(\log h^{-1})^{1/2}$, and define

$$\dot{\widehat{r}}_{\theta_o}(x) = \frac{1}{nh}\sum_{j=1}^n \dot{\Lambda}_{\theta_o}(Y_j)K_1\left(\frac{X_j - x}{h}\right), \quad \dot{\overline{r}}_{\theta_o}(x) = \mathbb{E}[\dot{\widehat{r}}_{\theta_o}(x)], \quad \overline{f}_X(x) = \mathbb{E}[\widehat{f}_X(x)],$$

where $\widehat{f}_X(x) = (nh)^{-1}\sum_{j=1}^n K_1(\frac{X_j - x}{h})$. Then,

$$\sup_{x\in\mathcal{X}_0}|\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)| \le \sup_{x\in\mathcal{X}_0}\left|\dot{\widehat{m}}_{\theta_o}(x) - \frac{\dot{\overline{r}}_{\theta_o}(x)}{\overline{f}_X(x)}\right| + \sup_{x\in\mathcal{X}_0}\frac{1}{\overline{f}_X(x)}\left|\dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x)\dot{m}_{\theta_o}(x)\right|. \tag{13}$$

Since $\mathbb{E}[\dot{\Lambda}^4_{\theta_o}(Y)|X = x] < \infty$ uniformly in $x \in \mathcal{X}$ by assumption (A7), a similar proof as was given for Theorem 2 in Einmahl and Mason (2005) ensures that

$$\sup_{x\in\mathcal{X}_0}\left|\dot{\widehat{m}}_{\theta_o}(x) - \frac{\dot{\overline{r}}_{\theta_o}(x)}{\overline{f}_X(x)}\right| = O_{\mathbb{P}}(c_n).$$

Consider now the second term of (**??**). Since $\mathbb{E}[\dot{\varepsilon}(\theta_o)|X] = 0$, where $\dot{\varepsilon}(\theta_o) = \frac{d}{d\theta}(\Lambda_\theta(Y) - m_\theta(X))|_{\theta=\theta_o}$, we have

$$\begin{aligned}
\dot{\overline{r}}_{\theta_o}(x) &= h^{-1}\mathbb{E}\left[\{\dot{m}_{\theta_o}(X) + \dot{\varepsilon}(\theta_o)\}K_1\left(\frac{X - x}{h}\right)\right]\\
&= h^{-1}\mathbb{E}\left[\dot{m}_{\theta_o}(X)K_1\left(\frac{X - x}{h}\right)\right]\\
&= \int \dot{m}_{\theta_o}(x + hv)K_1(v)f_X(x + hv)dv,
\end{aligned}$$

from which it follows that

$$\dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x)\dot{m}_{\theta_o}(x) = \int \left[\dot{m}_{\theta_o}(x+hv) - \dot{m}_{\theta_o}(x)\right] K_1(v) f_X(x+hv) dv.$$

Hence, a Taylor expansion applied to $\dot{m}_{\theta_o}(\cdot)$ yields

$$\sup_{x \in \mathcal{X}_0} \left|\dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x)\dot{m}_{\theta_o}(x)\right| = O(h^{q_1}) = O\left(c_n\right),$$

since $nh^{2q_1+1}(\log h^{-1})^{-1} = O(1)$ by (A2). This proves that the second term of (**??**) is $O(c_n)$, since it can be easily shown that $\overline{f}_X(x)$ is bounded away from 0 and infinity, uniformly in $x \in \mathcal{X}_0$, using (A3)(ii). □

References

[1] Ahmad, I. and Li, Q. (1997). Testing symmetry of an unknown density function by kernel method. *Journal of Nonparametric Statistics*, **7**, 279–293.

[2] Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, **28**, 549–567.

[3] Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge.

[4] Bickel, P.J. and Doksum, K. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association,* **76**, 296–311.

[5] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[6] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society - Series B*, **26**, 211–252.

[7] Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

[8] Chen, G., Lockhart, R.A. and Stephens, A. (2002). Box-Cox transformations in linear models: Large sample theory and tests of normality (with discussion). *Canadian Journal of Statistics,* **30**, 177–234.

[9] Cheng, F. (2005). Asymptotic distributions of error density and distribution function estimators in nonparametric regression. *Journal of Statistical Planning and Inference,* **128**, 327–349.

[10] Cheng, F. and Sun, S. (2008). A goodness-of-fit test of the errors in nonlinear autoregressive time series models. *Statistics and Probability Letters*, **78**, 50–59.

[11] Dette, H., Kusi-Appiah, S. and Neumeyer, N. (2002). Testing symmetry in nonparametric regression models. *Journal of Nonparametric Statistics*, **14**, 477–494.

[12] Efromovich, S. (2005). Estimation of the density of the regression errors. *Annals of Statistics,* **33**, 2194–2227.

[13] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics,* **33**, 1380–1403.

[14] Escanciano, J.C. and Jacho-Chavez, D. (2012). $\sqrt{n}$-uniformly consistent density estimation in nonparametric regression. *Journal of Econometrics*, **167**, 305–316.

[15] Fitzenberger, B., Wilke, R.A. and Zhang, X. (2010). Implementing Box-Cox quantile regression. *Econometric Reviews,* **29**, 158–181.

[16] Horowitz, J.L. (1998). *Semiparametric Methods in Economics.* Springer-Verlag, New York.

[17] Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Annals of Statistics,* **36**, 686–718.

[18] Müller, U.U., Schick, A. and Wefelmeyer, W. (2004). Estimating linear functionals of the error distribution in nonparametric regression. *Journal of Statistical Planning and Inference,* **119**, 75–93.

[19] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, 141—142.

[20] Neumeyer, N. and Dette, H. (2007). Testing for symmetric error distribution in nonparametric regression models. *Statistica Sinica*, **17**, 775–795.

[21] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis,* **101**, 1067–1078.

[22] Pinsker, M.S. (1980). Optimal filtering of a square integrable signal in Gaussian white noise. *Problems of Information Transmission*, **16**, 52–68.

[23] Sakia, R.M. (1992). The Box-Cox transformation technique: a review. *The Statistician,* **41**, 169–178.

[24] Samb, R. (2011). Nonparametric estimation of the density of regression errors. *Comptes Rendus de l'Académie des Sciences-Paris,* **Série I 349**, 1281-1285.

[25] Shin, Y. (2008). Semiparametric estimation of the Box-Cox transformation model. *Econometrics Journal,* **11**, 517–537.

[26] Vanhems, A. and Van Keilegom, I. (2011). Semiparametric transformation model with endogeneity: a control function approach. *Journal of Econometrics* (under revision).

[27] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā - Series A*, **26**, 359–372.

[28] Zellner, A. and Revankar, N.S. (1969). Generalized production functions. *Reviews of Economic Studies,* **36**, 241–250.

Postal addresses :

Benjamin Colling
Université catholique de Louvain
Institute of Statistics
Voie du Roman Pays 20
1348 Louvain-la-Neuve
Belgium

Cédric Heuchenne
HEC-Management School of the University of Liège,
Statistique appliquée à la gestion et à l'économie
Rue Louvrex 14, Bâtiment N1
4000 Liège
Belgium

Rawane Samb
Centre de recherche du CHUQ/CHUL
2705, Boulevard Laurier
G1V 4G2, QC
Québec
Canada

Ingrid Van Keilegom
Université catholique de Louvain
Institute of Statistics
Voie du Roman Pays 20
1348 Louvain-la-Neuve
Belgium